Supplemental Document for Adaptive User Modality-aware Preference Integration via Meta Learning Toward Multimodal Recommendation

Zhenchao Wu[1], Hongteng Xu[1], Xu Chen[1]
[1]Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China
{wuzhenchao, hongtengxu, xu.chen}@ruc.edu.cn

A. Pseudocode of Training Procedure

Algorithm 1: The training process of MeAIM

Data: user-item interactions
raw multimodal features of items
Hyperparameters: the batch size, the learning rate
the collaborative graph layers
the embedding dimension
α, β, λ, γ, τ, N
Result: the trained recommender model

Initialization
while *early stopping not reached* **do**
  foreach *batch* **do**
    Generate embeddings $\tilde{\mathbf{E}}_u^m$, $\tilde{\mathbf{E}}_i^m$, $m \in \{v, t, id\}$
    foreach *interaction* **do**
      Calculate $\hat{\omega}_{u^{m_u},i^{m_i}}$, $\hat{y}_{u^{m_u},i}$ according to Equ. (10)
      Calculate $\hat{\omega}_{u^{m_u},i}$, $\hat{y}_{u,i}$ according to Equ. (11)
    end
    Calculate $\mathcal{L}_{con}$, $\mathcal{L}_{KL}$, $\mathcal{L}_{bpr}$ according to Equ. (14), Equ. (16), and Equ. (19)
    Calculate $\mathcal{L}$ according to Equ. (20)
    Update parameters
  end
end

The pseudocode of the training process for MeAIM is shown in Algorithm 1. The MeAIM model first generates different modality representations for users and items, and then calculates the interaction probability between the user and the item. Afterward, the total loss is calculated and optimized to update the parameters of the MeAIM model.

B. Introduction of Baselines

We adopt two types of recommendation methods as baselines: (1) **General CF Methods**, including BPR [10], LightGCN [11], and LayerGCN [15]. (2) **Multimodal Methods**, including VBPR [1], MMGCN [2], GRCN [3], LATTICE [4], SLMRec [9], BM3 [12], MMSSL [8], FREEDOM [5], DA-MRS [6], and LGMRec [7]. Below, we introduce these baselines one by one.

Two general collaborative filtering methods make recommendations for users according to their interaction with items:

- **BPR** is a personalized ranking method, whose optimization criterion and learning algorithm are the maximum posterior estimator and stochastic gradient descent, respectively.
- **LightGCN** first learns user and item embeddings only with neighborhood aggregation at each layer, and then generates final embeddings by the weighted sum of the learned embeddings at all layers.
- **LayerGCN** refines layer representations to alleviate over-smoothing during information propagation and node updating of GCN, and adopts a degree-sensitive probability to prune the noise edges.

In addition to interaction information, multimodal recommender systems leverage multimodal features to improve representation learning for better recommendation performances:

- **VBPR** incorporates the visual features of items into Matrix Factorization, which concatenates the visual embeddings with ID embeddings as final item embeddings.
- **MMGCN** constructs a user-item graph in each modality, and utilizes the user-item interactions to improve the representation learning of nodes.
- **GRCN** designs a refining layer to identify and prune the false-positive edges in the user-item bipartite graph. The user preferences are learned by a GCN layer on the refined graph.
- **LATTICE** first builds the item-item graph for each modality, which is combined to obtain the multimodal item relationship graph. This strategy could enhance item representations.
- **SLMRec** applies self-supervised learning to different views of each item, which helps capture the potential relations among modalities for better representation learning.
- **BM3** designs a multi-modal contrastive loss to learn the representations of users and items without auxiliary graphs and negative samples.
- **MMSSL** utilizes self-supervised learning to learn the user modality-aware preference, and employs cross-modal contrastive learning to preserve the inter-modal semantic commonality and the user preference diversity.
- **FREEDOM** freezes the item-item graph during the training process for a lower cost of computation and memory,

and denoises the user-item interaction graph with a degree-sensitive edge pruning method.
- **DA-MRS** deals with the noise existing in the multimodal content and user feedback, and achieves the multi-modal alignment guided by user preference and graded item relations.
- **LGMRec** learns local and global embeddings of users and items according to local topological information and hypergraph dependencies, respectively.

C. Implementation Details of Methods

We run relevant experiments on a Linux platform with NVIDIA A40 GPU cards (48 GB) under the environment Pytorch 1.8.2. MeAIM and some baselines are implemented based on the MMRec framework[1]. In experiments, we utilize the preprocessed datasets from LGMRec [7] to evaluate MeAIM and baselines. Therefore, we directly adopt the experimental results of some baselines from the paper [7], including [1]–[4], [7], [9]–[12]. Here, we only expound on the implementation details of MeAIM and the baselines that we reproduced. Following the work [7], we fix the embedding dimension of users and items to 64, the batch size to 2048, and the learning rate to 0.001 in all models. In all GNN-based methods, the collaborative ID graph layers are set to 2. Based on these settings, MeAIM and some baselines may not achieve optimal results, but this strategy could ensure a fair comparison. For other key hyperparameters of different methods, we adopt grid search to seek optimal results.

First, we describe the implementation details of MeAIM. For collaborative modality graph layers, we set visual graph layers and textual graph layers to 2 and 3 on Baby, 2 and 4 on Sports, and 3 and 3 on Clothing. For the item-item semantic graphs, the number of similar items N is fixed to 10, and the weight of the visual modality α is set to 0.1 on all datasets. In the meta-weight net, we set the dropout rate to 0.5 on Baby and Clothing, and 0.7 on Sports. Besides, the weight coefficients β, λ, and γ are set to 0.001, 100, and 0.5 on Baby, 0.001, 10000, and 0.5 on Sports, and 0.0001, 100, and 1.0 on Clothing. In the contrastive learning formula, the temperature factor τ is set to 0.2. The early stopping is set to 20 on Baby while 30 on Sports and Clothing. The total epochs are set to 1000.

Next, we introduce the settings of key parameters for other baselines. For LayerGCN, we set the regularizer weight to 0.01 on the three datasets, and set the dropout rate to 0.2 on Baby and Clothing and 0.0 on Sports. For MMSSL, we set augmentation factors ζ, λ_1 in Adversarial SSL, and the temperature parameter to 100, 1, and 0.5 on the three datasets, respectively. For FREEDOM, we set the layers of the item-item graph and the number of neighbors to 1 and 10 on the three datasets, respectively. Besides, we set the ratio of the degree-sensitive edge pruning and loss weight λ are 0.7 and 0.01 on Baby, 0.9 and 0.001 on Sports, and 0.9 and 0.0001 on Clothing. For DA-MRS, we adopt the same parameter settings with FREEDOM on the layers of the user-item bipartite graph and item-item graph, and the number of neighbors. In addition, we set the neighbor weight and the KL weight to 0.001 and 1 on the three datasets. We also set $\alpha = \beta = 1.5$ and $\gamma = 2$ on Baby, $\alpha = \beta = 3$ and $\gamma = 1$ on Sports, and $\alpha = \beta = 1.5$ and $\gamma = 1$ on Clothing.

[1]https://github.com/enoche/MMRec

TABLE I: Experimental results on benchmark datasets.

Model	Baby		Sports		Clothing	
	R@10	N@10	R@10	N@10	R@10	N@10
BPR	0.0379	0.0202	0.0452	0.0252	0.0211	0.0118
LightGCN	0.0464	0.0251	0.0553	0.0307	0.0331	0.0181
LayerGCN	0.0507	0.0273	0.0601	0.0330	0.0364	0.0198
VBPR	0.0424	0.0223	0.0556	0.0301	0.0281	0.0158
MMGCN	0.0398	0.0211	0.0382	0.0200	0.0229	0.0118
GRCN	0.0531	0.0291	0.0600	0.0324	0.0431	0.0230
LATTICE	0.0536	0.0287	0.0618	0.0337	0.0459	0.0253
SLMRec	0.0540	0.0296	0.0676	0.0374	0.0452	0.0247
BM3	0.0538	0.0301	0.0659	0.0354	0.0450	0.0243
MMSSL	0.0620	0.0353	0.0683	0.0394	0.0505	0.0278
FREEDOM	0.0643	0.0340	0.0716	0.0383	0.0630	0.0336
DA-MRS	0.0623	0.0331	0.0704	0.0381	0.0622	0.0333
LGMRec	0.0644	0.0349	0.0720	0.0390	0.0555	0.0302
MeAIM	**0.0662**	**0.0356**	**0.0776**	**0.0425**	**0.0660**	**0.0355**
Improv. (%)	2.80	0.85	7.78	7.87	4.76	5.65

D. Overall Performance

In this section, we supplement the experimental results of MeAIM and all baselines on Recall@10 and NDCG@10 in TABLE I. This table shows that MeAIM consistently outperforms all baselines on the three datasets. Therefore, the experimental results further validate the rationality of MeAIM in design.

E. Parameter Analysis

For MeAIM, we evaluate the sensitivity of several key hyperparameters across different datasets. These hyperparameters include the collaborative graph layers about the visual and textual modality, the contrastive learning loss weight β, the KL loss weight λ, the weight coefficient γ in Eq (**??**), and the dropout rate ρ in the meta-weight net.

Given a fair comparison, we have fixed the collaborative ID graph layers to 2. Under this setting, we study the impact of collaborative modality graph layers. In this experiment, we set the visual and textual graph layers from 1 to 4, and the experimental results are shown in Fig. 1. For different datasets, the optimal results are obtained when the modality graph layers are different. Specifically, the visual and textual graph layers are set to 2 and 3 on Baby, 2 and 4 on Sports, and 3 and 3 on Clothing when the results are best. The phenomenon reflects the necessity of adjusting the modality graph layers when applying the recommender model.

The contrastive learning task (CL) and KL divergence constraint (KL) are adopted to enhance the modality-aware representations. Here, we investigate how the CL loss weight β and the KL loss weight λ affect recommendation performances. Fig. 2 shows the distribution of experimental results under different CL loss weight settings and KL loss weight settings. We observe that the corresponding weights for the best results are different on the three datasets. The optimal results are obtained when the parameters β and λ are set to 0.001 and 100 on Baby, 0.001 and 10000 on Sports, and

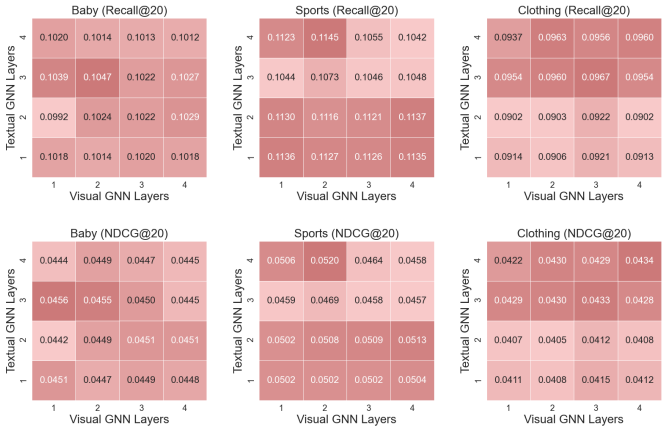

Fig. 1: Impact of collaborative visual and textual graph layers.

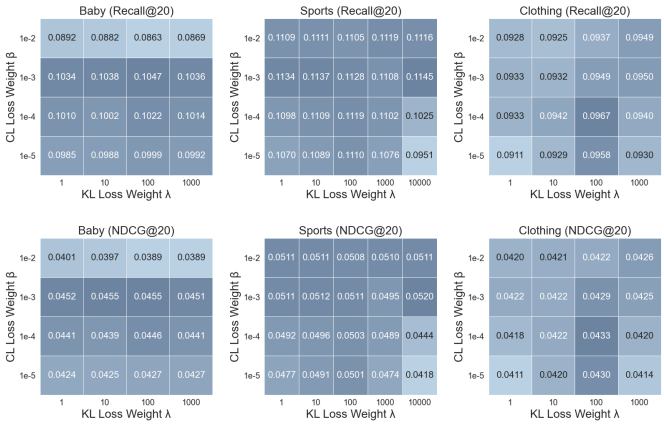

Fig. 2: Impact of the KL loss weight λ and the CL loss weight β.

0.0001 and 100 on Clothing. Notably, the ablation experiments have proved the effectiveness of the two terms.

The weight coefficient γ and the dropout rate ρ in the meta-weight net are two key hyperparameters, which largely affect recommendation performances. The parameter γ controls the influence of further incorporating the BPR loss about user ID. The larger the value γ is, the less the corresponding influence is. Fig. 3 presents the experimental results under different settings of the γ value. The recommendation performances are optimal when the value γ is set to 0.5 on Baby and Sports, and 1.0 on Clothing. In addition, Fig. 4 exhibits the experimental results on the three datasets as the value ρ varies from 0 to 0.9. In terms of final results, we set the ρ value to 0.5 on Baby, 0.7 on Sports, and 0.5 on Clothing, respectively. In Fig. 4, we observe that these settings may not correspond to the optimal results on the metrics $Recall@20$ and $NDCG@20$. But it is a compromise when taking the results on $Recall@10$ and $NDCG@10$ into account.

F. Visualization Analysis

In this paper, we design a cross-modality contrastive learning task and the KL divergence constraint to enhance modality-aware representations. To validate the effectiveness of the two

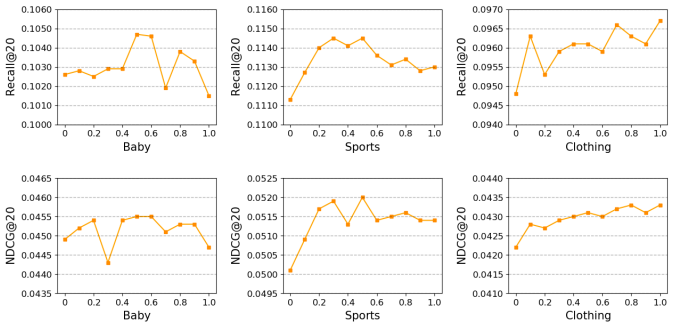

Fig. 3: Impact of the weight coefficient γ.

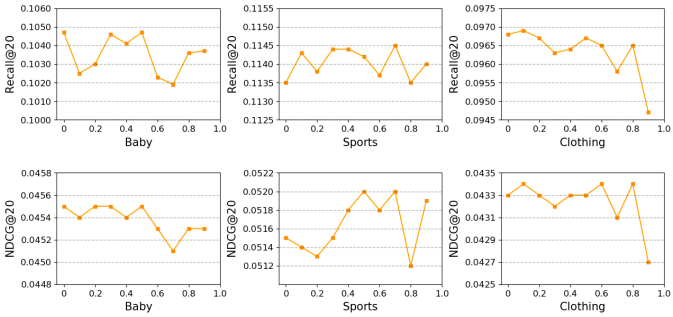

Fig. 4: Impact of the dropout rate ρ in meta-weight net.

components, we visualize the distribution of modality-aware representations of users and items. It has been proved that the textual modality contains more informative content than the visual modality in the ablation experiments. Therefore, we randomly select 500 users and 500 items from Sports, and then employ t-SNE [13] to map the textual representations to a two-dimensional space.

Fig. 5 and Fig. 6 present the distribution of 2D textual representations about users and items with and without the two components, respectively. At the bottom of each picture, we plot the density estimation of arctan(x, y), where (x, y) denotes the Cartesian coordinates of 2D representations. Here, we define "MeAIM-CK" as the recommender model when the two components are removed from MeAIM. Compared to MeAIM-CK, the distribution of 2D textual features in MeAIM is more dispersed and uniform, which indicates that more discriminative modality representations can be generated when contrastive learning and the KL divergence constraint are utilized. Previous work [14] has indicated that the more uniform the distribution of representations is, the higher-quality representations are. Therefore, the two components can help enhance modality representations and improve recommendation performance. Notably, the recommendation performance will degrade when removing any component in the ablation study. This phenomenon is consistent with the uniform distribution of modality representations.

REFERENCES

[1] R. He and J. McAuley, "VBPR: Visual Bayesian Personalized Ranking from Implicit Feedback," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 2016, pp. 144-150.

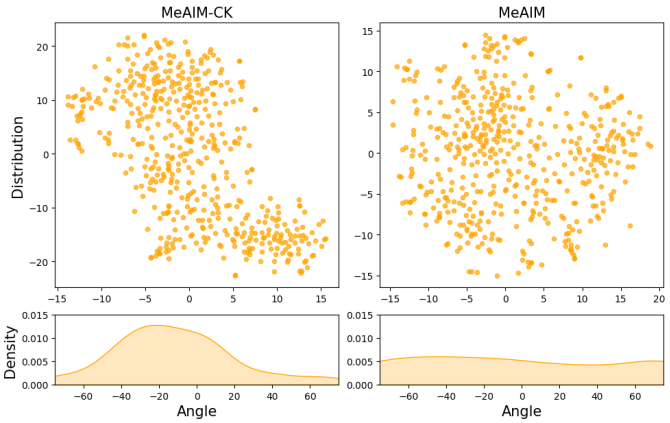

Fig. 5: The distribution of user textual representations on Sports.

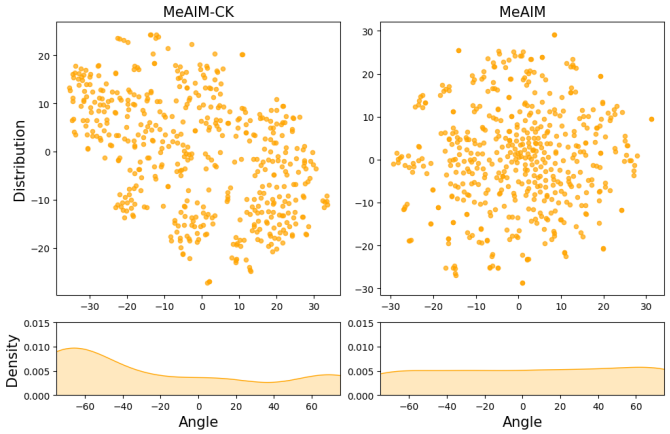

Fig. 6: The distribution of item textual representations on Sports.

[2] Y. Wei, X. Wang, L. Nie, X. He, R. Hong, and T. Chua, "MMGCN: Multi-modal Graph Convolution Network for Personalized Recommendation of Micro-video," in *Proceedings of the 27th ACM international conference on multimedia*, 2019, pp. 1437–1445.

[3] Y. Wei, X. Wang, L. Nie, X. He, and T. Chua, "Graph-Refined Convolutional Network for Multimedia Recommendation with Implicit Feedback," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 3541-3549.

[4] J. Zhang, Y. Zhu, Q. Liu, S. Wu, S. Wang, and L. Wang, "Mining Latent Structures for Multimedia Recommendation," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 3872–3880.

[5] X. Zhou and Z. Shen, "A Tale of Two Graphs: Freezing and Denoising Graph Structures for Multimodal Recommendation," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 935–943.

[6] G. Xv, X. Li, R. Xie, C. Lin, C. Liu, F. Xia, Z. Kang, and L. Lin, "Improving Multi-modal Recommender Systems by Denoising and Aligning Multi-modal Content and User Feedback," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024, pp. 3645–3656.

[7] Z. Guo, J. Li, G. Li, C. Wang, S. Shi, and B. Ruan, "LGMRec: Local and Global Graph Learning for Multimodal Recommendation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, pp. 8454-8462.

[8] W. Wei, C. Huang, L. Xia, and C. Zhang, "Multi-Modal Self-Supervised Learning for Recommendation," in *Proceedings of the ACM Web Conference*, 2023, pp. 790–800.

[9] Z. Tao, X. Liu, Y. Xia, X. Wang, L. Yang, X. Huang, and T. Chua, "Self-Supervised Learning for Multimedia Recommendation," in *IEEE Transactions on Multimedia*, 2023, pp. 5107-5116.

[10] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, "BPR: Bayesian personalized ranking from implicit feedback," in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, 2009, pp. 452-461.

[11] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, and M. Wang, "LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020, pp. 639–648.

[12] X. Zhou, H. Zhou, Y. Liu, Z. Zeng, C. Miao, P. Wang, Y. You, and F. Jiang, "Bootstrap Latent Representations for Multi-modal Recommendation," in *Proceedings of the ACM Web Conference*, 2023, pp. 845–854.

[13] L. V. d. Maaten and G. Hinton, "Visualizing data using t-SNE," in *Journal of machine learning research*, 2008.

[14] C. Wang, Y. Yu, W. Ma, M. Zhang, C. Chen, Y. Liu, and S. Ma, "Towards representation alignment and uniformity in collaborative filtering," in *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, 2022, pp. 1816–1825.

[15] X. Zhou, D. Lin, Y. Liu, and C. Miao, "Layer-refined Graph Convolutional Networks for Recommendation," in *IEEE 39th International Conference on Data Engineering (ICDE)*, 2023, pp. 1247–1259.